DATAESTUR

Experimental use case: Generating sentiment-based indices from digital listening data

Methodology and application

Publication date: October 2024

Table of Contents

List of illustrations

List of tables

I. Introduction

The purpose of this document is to explain the methodology developed and used by SEGITTUR to build an sentiment-based index based. A general index of travellers' opinions on tourism in Spain has been created. This document provides an example of a methodology for preparing an index, which can be applied to various areas, such as a satisfaction or reputation index.

The index is built from a database of *digital listening*. This source collects mentions from users in various online channels for a previously defined topic and classifies them by *sentiment* (positive, negative or neutral assessment of the text).

Essentially, the work carried out consists of processing *digital listening* data through a programming language, in this case the free *R software*, thus creating an index. It is possible to do this same thing with any other type of software.

II. Methodology

An index is an indicator that allows for the assessment of performance in a specific area over a period of time. The use of indexes as an evaluation tool for a product or service allows measuring satisfaction over time, which is associated with the resolution of a need. This is a complex field where, according to studies, different variables have an influence. The main satisfaction models are linked to opinions regarding expectations, perception of quality, intention to recommend or continue using, tolerance to variations in price or quality, management of complaints or incidents, competitive advantages of the product, brand image or emotions aroused.

Based on these opinions, it is possible to build an index. However, it is important to determine the model for gathering opinions. A common method is the survey or panel system. But with the widespread use of the Internet and Social Media, the possibility of

gathering opinions through web posts has enabled the generation databases therefrom. In this document, this is called *digital* listening. This work is based on this last model. *Digital listening* allows opinions to be categorised and a *sentiment* to be assigned to them, as ascertained by assessing the opinion text and categorising it according to whether it expresses a positive *sentiment*, a negative *sentiment* or a neutral *sentiment*.

The practical case developed is described below, setting out the necessary elements (inventory) and the calculation model.

1. Inventory of resources required for the preparation of the Index.

First, an inventory of elements is suggested for application of the methodology in this document.

- Database with user opinions.
- Elements for data cleansing and variable creation.
- Preparation of a table defining strata and their weighting.

1.1. Database with user opinions.

As a starting point, there must be variables to segment the sample and thus correct possible sampling errors.

To determine opinions, we must define the topic on which information is to be obtained, the time axis of the data and the periodicity. Each opinion has an identification number (ID) and, in this case, it will be assigned the channel where it was posted, the language of the text, the *sentiment* conveyed in the opinion, and the subject matter of the opinion. These variables must be representative. For this methodology, the main variables used and their possible values are detailed in the following illustration:

Figure 1 Key variables analysed and possible components

Channel	Language	Sentiment	Topic
•Blogs •Forums •Facebook •News •Twitter •Webs	•Castilian •Catalan •Galician •Basque	•Positive •Negative •Neutral	•Journey •Destination •Hotel •Transport •Environment

Source: by the authors.

Each of the values within each variable can be coded to facilitate the work. In this methodology, a number has been assigned to each channel value, to each possible language and to each *sentiment*, and this coding has been used to construct the strata described in the following section.

1.2. Elements for data cleansing and variable creation.

Raw data may require some processing. That is, it may be necessary to review nulls, duplicates, and invalid fields. Therefore, to facilitate data quality and index acquisition the following elements should be included:

- ***Creating a list with sentiment scoring***. In this case, the suggestion is to use 100 if it is positive, 50 if it is neutral and 0 if it is negative.
- ***Dictionary and coding/keys of variables***. It can be in one of the previous tables or directly in the code script.
- ***List of valid fields for languages, channels and sentiments***. Use the already defined numerical key and include it in the code script that will process the data. This is how IDs with invalid values are removed.

1.3. Framework for estimation.

The estimation can be carried out in different ways. There must be a *"universe"* or reference values to extrapolate the sample data to the population and build the indicator. In our case, we have chosen to segment the sample into different strata, and made an estimation based on the representativeness of these strata.

For this reason, data classification groups (*"strata"*) have been developed for the estimation, establishing a minimum number of data for a stratum. Data weighting is established in the same way. These concepts are further explained below.

- ***Strata***

Strata are homogeneous groups within the sample, so each element within the same stratum has its own characteristics that differentiate it from those in other groups. Their development is necessary to assign different importance to certain groups in the sample.

This methodology builds the strata of the Channel and Language combination. For the application of *digital listening*, three levels of strata have been defined. These allow the stratum to be weighted as long as there is a representative sample thereof and, as the level advances, the model becomes simpler with fewer combinations of variables. For the case studied, level 1 has 24 strata derived from the combination of each language with each channel. Level 2 has 8 strata, due to the grouping of some channels; and in Level 3 there are 4 strata, due to the grouping of Level 2 channels, and the grouping of languages. The objective of these groupings is to guarantee representative samples in each stratum and to simplify the model.

Figure 2 Levels of strata

Strata Level 1	Strata Level 2	Strata Level 3
•Each channel •Each language	•Channels grouped into 2: Social Media and Internet •Each language	•Channels grouped into 2: Social Media and Internet. •Languages grouped into 2: Spanish and co-official languages.

Source: by the authors.

Table 1 *Summary of strata* shows the construction of strata for the three levels. It is possible to identify which channel and which language each stratum is composed of at levels 1, 2 or 3, as well as which coding refers to that stratum and what its description is.

For example, at level 1, channel 4 is *News* and language 2 is *Catalan*. They make up stratum 42, called *News* and *Catalan*. At level 2, some channels are reduced. For the

same example, the stratum becomes 12 described as *Internet* and *Catalan*, since it includes the *Internet* channels (grouping several channels) and the *Catalan* language. At level 3, languages are grouped together. In this case, code 12 is stratum 3, called *Internet* and *Main co-official languages*.

To this table, the weight acquired by each stratum will be added according to the importance criteria explained below in the section *Weighting or weights*.

Table 1 Summary of strata

CHANNEL_STRATUM_1	LANGUAGE_STRATUM_1	STRATUM_1	DESC_CHANNEL_STRATUM_1	DESC_LANGUAGE_STRATUM_1	DESC_STRATUM_1	CHANNEL_STRATUM_2	LANGUAGE_STRATUM_2	STRATUM_2	DESC_CHANNEL_STRATUM_2	DESC_LANGUAGE_STRATUM_2	DESC_STRATUM_2	CHANNEL_STRATUM_3	LANGUAGE_STRATUM_3	STRATUM_3	DESC_CHANNEL_STRATUM_3	DESC_LANGUAGE_STRATUM_3	DESC_STRATUM_3
1	1	11	Blogs	Castilian	Blogs and Castilian	1	1	11	Internet	Castilian	Internet and Castilian	1	1	11	Internet	Castilian	Internet and Castilian
1	2	12	Blogs	Catalan	Blogs and Catalan	1	2	12	Internet	Catalan	Internet and Catalan	1	2	12	Internet	Main co-official	Internet and Main co-official
1	3	13	Blogs	Basque	Blogs and Basque	1	3	13	Internet	Basque	Internet and Basque	1	2	12	Internet	Main co-official	Internet and Main co-official
1	4	14	Blogs	Galician	Blogs and Galician	1	4	14	Internet	Galician	Internet and Galician	1	2	12	Internet	Main co-official	Internet and Main co-official
2	1	21	Discussions	Castilian	Discussions and Castilian	1	1	11	Internet	Castilian	Internet and Castilian	1	1	11	Internet	Castilian	Internet and Castilian
2	2	22	Discussions	Catalan	Discussions and Catalan	1	2	12	Internet	Catalan	Internet and Catalan	1	2	12	Internet	Main co-official	Internet and Main co-official
2	3	23	Discussions	Basque	Discussions and Basque	1	3	13	Internet	Basque	Internet and Basque	1	2	12	Internet	Main co-official	Internet and Main co-official
2	4	24	Discussions	Galician	Discussions and Galician	1	4	14	Internet	Galician	Internet and Galician	1	2	12	Internet	Main co-official	Internet and Main co-official
3	1	31	Facebook	Castilian	Facebook and Castilian	2	1	21	Social Media	Castilian	Social Media and Castilian	2	1	21	Social Media	Castilian	Social Media and Castilian
3	2	32	Facebook	Catalan	Facebook and Catalan	2	2	22	Social Media	Catalan	Social Media and Catalan	2	2	22	Social Media	Main co-official	Social Media and Main co-official
3	3	33	Facebook	Basque	Facebook and Basque	2	3	23	Social Media	Basque	Social Media and Basque	2	2	22	Social Media	Main co-official	Social Media and Main co-official
3	4	34	Facebook	Galician	Facebook and Galician	2	4	24	Social Media	Galician	Social Media and Galician	2	2	22	Social Media	Main co-official	Social Media and Main co-official
4	1	41	News	Castilian	News and Castilian	1	1	11	Internet	Castilian	Internet and Castilian	1	1	11	Internet	Castilian	Internet and Castilian
4	2	42	News	Catalan	News and Catalan	1	2	12	Internet	Catalan	Internet and Catalan	1	2	12	Internet	Main co-official	Internet and Main co-official
4	3	43	News	Basque	News and Basque	1	3	13	Internet	Basque	Internet and Basque	1	2	12	Internet	Main co-official	Internet and Main co-official
4	4	44	News	Galician	News and Galician	1	4	14	Internet	Galician	Internet and Galician	1	2	12	Internet	Main co-official	Internet and Main co-official
5	1	51	Twitter	Castilian	Twitter and Castilian	2	1	21	Social Media	Castilian	Social Media and Castilian	2	1	21	Social Media	Castilian	Social Media and Castilian
5	2	52	Twitter	Catalan	Twitter and Catalan	2	2	22	Social Media	Catalan	Social Media and Catalan	2	2	22	Social Media	Main co-official	Social Media and Main co-official
5	3	53	Twitter	Basque	Twitter and Basque	2	3	23	Social Media	Basque	Social Media and Basque	2	2	22	Social Media	Main co-official	Social Media and Main co-official
5	4	54	Twitter	Galician	Twitter and Galician	2	4	24	Social Media	Galician	Social Media and Galician	2	2	22	Social Media	Main co-official	Social Media and Main co-official
6	1	61	Webs	Castilian	Webs and Castilian	1	1	11	Internet	Castilian	Internet and Castilian	1	1	11	Internet	Castilian	Internet and Castilian
6	2	62	Webs	Catalan	Webs and Catalan	1	2	12	Internet	Catalan	Internet and Catalan	1	2	12	Internet	Main co-official	Internet and Main co-official
6	3	63	Webs	Basque	Webs and Basque	1	3	13	Internet	Basque	Internet and Basque	1	2	12	Internet	Main co-official	Internet and Main co-official
6	4	64	Webs	Galician	Webs and Galician	1	4	14	Internet	Galician	Internet and Galician	1	2	12	Internet	Main co-official	Internet and Main co-official

Source: by the authors.

- **Weighting or weights.**

This methodology proposes evaluating each positive, negative or neutral mention based on a weight depending on which stratum it belongs to. This weight has been developed with a double weighting that is intended to remain fixed over time:

- **Weight A:** the weight of the stratum in the sample. It is assumed that less frequent strata could be considered less significant and, therefore, their *sentiment* will score less than that of strata with greater weight.
  This weight is not included in the strata summary table, since it is calculated in the database itself based on the volume of data in each stratum. Its definition has been included in this section to improve the interpretation of double weighting; its calculation will be further explained in the section *Applied calculation process*.

- **Weight B:** the weight of the stratum according to its representativeness in the population or similar. It arises from the idea that there may be different

representativeness depending on whether the opinion is from a specific language or channel.

As an example, the *digital listening* sample received could have a similar volume of observations in Galician and Castilian. According to INE[1] statistics on the frequency of language use, Galician is spoken by 4.11% of people living in Spain and, since Castilian is a language with a higher percentage of speakers, it is suggested that the weights of each language be adjusted by weighting them. This helps to achieve a representativeness that is as close to reality as possible. This also applies to the channel: if the use of the Internet to read news is significantly higher than that of posting in blogs, we suggest taking this into account by adding a weighting to each channel.

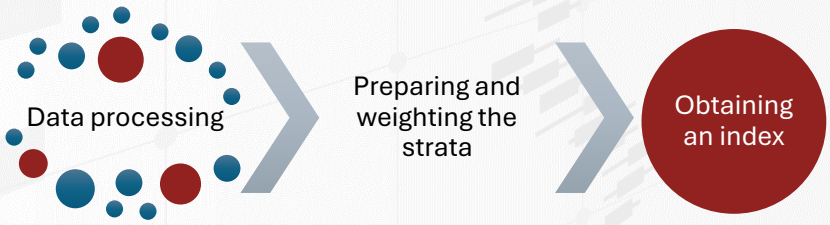
In conclusion, weighting by language and weighting by channel result in a single weighting for each stratum by multiplying both – called weight B and included in the strata summary table.

2. Applied calculation process.

Once the inventory is defined, work with the database will begin. However, as its analysis progresses, it might provide more specific information about the strata, and we can refer back to the summary table and review the approach to the different strata levels. This review can be repeated as the process progresses to adjust it.

The work carried out is outlined below, consisting of three phases represented in *Illustration 3*:

Figure 3 Process diagram for developing a global sentiment-based index



[1] Survey: National Institute of Statistics. (2021). Survey of Essential Characteristics of the Population and Housing. Madrid, Spain: INE.

Source: by the authors.

2.1. Data processing: cleansing and coding.

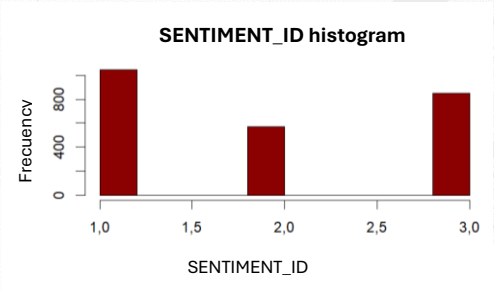
First of all, we should define process periodicity. We propose to generate the index monthly, so a time variable will be included, "*Year/month*".

An assessment of the digital *listening* database is recommended. Specifically, identify possible repeated rows, null or empty data. Similarly, debug invalid values (codes not included in the list of valid channels, languages and sentiments). These invalid or null values that we are going to eliminate from the database should be included in a control table.

To facilitate understanding, we suggest including the semantic meaning of each variable language, channel, *sentiment* and stratum if it is not found in the database.

It is interesting to study the distribution of data by channel, language or sentiment. You can use a histogram or a list that groups unique values together by categories. A better understanding of the sample will have a positive impact on the analysis of results and their adjustment. The quality of the sample will be key to the quality of the calculated index.

Figure 4 Sentiment histogram for studying data distribution



Source: by the authors.

2.2. Preparing and weighting the strata.

The first step is to incorporate the strata table already mentioned in the *Inventory* section, which included each stratum's coding, semantic explanation and weight B (generated from statistical data from reliable sources).

Once it is imported into the code programme (R), the *Weight table* will be created. In order to ensure completeness, the volume of data by stratum and level will be incorporated from the *digital listening* database. This enables us to see how each stratum is distributed at each

level, detecting whether a stratum is unrepresentative. For this case, a premise has been established: if a stratum has less than 10 values, the merger of strata to the next stratum level will be applied. Let's see an example with the support of *Table 2*, which represents the volumes obtained from each stratum at different levels.

- At the stratum level 1 (column STRATUM_1) for stratum 11 (which the summary table shows a Blogs in Castilian), there are 3 values in the database. In this case, since there are less than 10 values, their stratum level 2 will be used (column STRATUM_2), which adds up to 297 observations.
- This merger involves redoing all level 1 strata that share stratum 11 at level 2 so as not to increase the number of data. The example shows how this rule has been applied, and there are no repeated rows from stratum 11 at level 2.
- This same process can be applied at level 2 with a merger at level 3. In the example in *Table 2*, there are enough values at level 2 in all strata (more than 9), so level 2 will be used in both Sum and Weight B to calculate the Final Weight.

Table 2 Adjusted strata table with sum and weight B (partial view)

	STRATUM_1	STRATUM_2	STRATUM_3	Stratum_1_sum	Stratum_2_sum	Stratum_3_sum	WEIGHT_B_STRATUM_1	WEIGHT_B_STRATUM_2	WEIGHT_B_STRATUM_3
1	11	11	11	3	297	297	0.144468531	0.53338466	0.53338466
2	12	12	12	44	504	961	0.029707784	0.10968255	0.15234758
3	13	13	12	33	118	961	0.003931238	0.01451432	0.15234758
4	14	14	12	3	339	961	0.007624692	0.02815072	0.15234758
5	31	21	21	2	147	147	0.146446334	0.24444760	0.24444760
6	32	22	22	0	488	1039	0.030114490	0.05026698	0.06982016
7	33	23	22	1	291	1039	0.003985057	0.00665184	0.06982016
8	34	24	22	0	260	1039	0.007729075	0.01290134	0.06982016

Source: by the authors.

Once we have restructured the strata, we suggest calculating Weight A and Final weight. As an example, the formula for *stratum 11*:

Figure 5 Calculation of Weight A and Final Weight

$$Weigth\ A_{\ stratum\ 11} = \frac{Sum\ of\ variables_{\ stratum\ 11}}{Sum\ of\ variables_{\ total\ sample}}$$

$$Final\ Weight_{\ stratum\ 11} = \frac{Weight\ B_{\ stratum\ 11}}{Sum\ of\ variables_{\ stratum\ 11}} * Sum\ of\ variables_{\ total\ sample}$$

Source: by the authors.

We suggest incorporating a control table to check that the sum of weight A and the sum of weight B must be equal to 1 respectively. Also, we consider it interesting to include a check for the Final weight, which should be the sum of the Final weight divided by the total number of observations equal to 1. This table could be the same one for cleaning duplicates and nulls.

2.3. Obtaining a global index.

From this stage onwards, we can work directly on the *digital* listening database. To do this, the Final weight is incorporated into it by linking according to the stratum indicated by each row or ID. We would still need to calculate the score for the *sentiment*, which we have defined in the inventory section: 100 for positive comments, 50 for neutral comments and 0 for negative comments. The ID Score is obtained by multiplying the Final weight by the *sentiment* score of each ID.

Table 3 Database (partial view) with Final weighted, sentiment score and ID

LANGUAGE_DESC	SENTIMENT_DESC	GENDER_DESC	STRATUM_FINAL_WEIGHTED	SENTIMENT_SCORE	ID_SCORE
Castilian	Positive	Unknown	4,064149	100	0,16629088
Castilian	Positive	Unknown	4,389199	100	0,17959079
Castilian	Positive	Unknown	4,389199	100	0,17959079
Castilian	Positive	Female	4,389199	100	0,17959079
Castilian	Neutral	Female	4,389199	50	0,08979540

Source: by the authors.

The sum of the ID Score column will give us the indicator value for the assessed period. We can also assess the contribution of each topic to the Global Index, by adding the ID Score per topic. *Table 4* shows the Global Index for the selected period, which is 60.61.

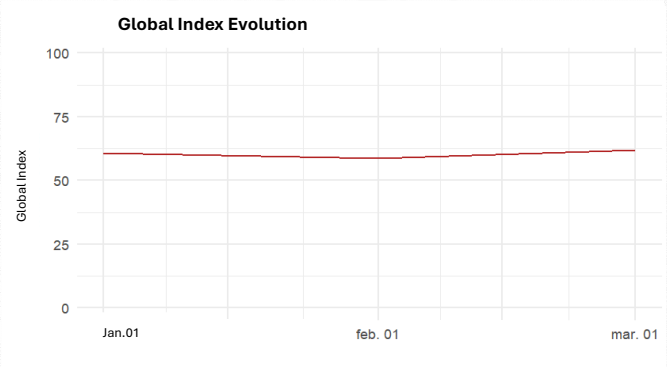
Table 4 Global Index obtained from the sum of ID Scores for a period

	Year.Month	Global.Index
1	202401	60,61706

Source: by the authors.

We suggest exporting the result to an Excel file to record the index data for each period evaluated in order to have the monthly indexes recorded.

Figure 6 Global Index evolution chart



Source: by the authors.

This same methodology could be applied to generate *digital* listening indexes more specific to concrete topics such as accommodation, transport, environment or destination. To do this, we suggest applying an initial filter to the data to get opinions that address the topic to be evaluated.

III. Conclusions

To conclude, with this case, it has been possible to construct a unique Index or value from a database containing information relating to opinions and sentiment. There are different ways to do this, and the case presented is an example therefor.

The construction of the index requires an organised approach to data segmentation, its configuration as strata and the definition of weightings for the strata in order to balance the representativeness of the different categories. Equally relevant is prior data processing or cleansing.

Finally, there are various uses for building an index with this methodology or any other, since it is possible to create an index using other procedures and softwares. There are many more applications and developments that can be carried out and an index can be built through specific segmentation by types of topics, keywords, demographic groups or other variables that form part of the database. All of this allows for the development of a more complete index system that can assess overall satisfaction in the same way, by areas, groups or topics, helping to interpret the general index.

NOTES

Methodological notes

This document aims to generate a methodology for the creation of indexes. It is possible to develop it from a different approach or using other techniques.

The free R program was used for data processing and index construction, although other programs can be used.

For specific questions or suggestions, please contact us via info.sit@segittur.es.

Publication date: October 2024

DATAESTUR